

<https://helda.helsinki.fi>

pö Bibliographic Data Science and the History of the

Lahti, Leo

2019

Lahti , L , Marjanen , J P , Roivainen , H H M & Tolonen , M S 2019 , ' Bibliographic Data
pö Science and the History of the Book (c. 1500 1800) ' , Cataloging & Cl
, vol. 57 , no. 1 , pp. 5-23 . <https://doi.org/10.1080/01639374.2018.1543747>

<http://hdl.handle.net/10138/301713>

<https://doi.org/10.1080/01639374.2018.1543747>

cc_by

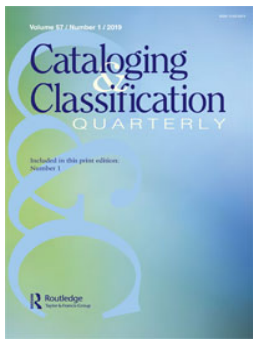
publishedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.



Bibliographic Data Science and the History of the Book (c. 1500–1800)

Leo Lahti, Jani Marjanen, Hege Roivainen & Mikko Tolonen

To cite this article: Leo Lahti, Jani Marjanen, Hege Roivainen & Mikko Tolonen (2019) Bibliographic Data Science and the History of the Book (c. 1500–1800), *Cataloging & Classification Quarterly*, 57:1, 5-23, DOI: [10.1080/01639374.2018.1543747](https://doi.org/10.1080/01639374.2018.1543747)

To link to this article: <https://doi.org/10.1080/01639374.2018.1543747>



Published with license by Taylor & Francis©
Leo Lahti, Jani Marjanen, Hege Roivainen
and Mikko Tolonen.© Leo Lahti, Jani
Marjanen, Hege Roivainen and Mikko
Tolonen.



View supplementary material [↗](#)



Published online: 07 Jan 2019.



Submit your article to this journal [↗](#)



Article views: 1026



View Crossmark data [↗](#)

Bibliographic Data Science and the History of the Book (c. 1500–1800)

Leo Lahti^a , Jani Marjanen^b , Hege Roivainen^b , and Mikko Tolonen^b 

^aDepartment of Mathematics and Statistics, University of Turku, Finland; ^bHelsinki Computational History Group, Department of Digital Humanities, University of Helsinki, Finland

ABSTRACT

National bibliographies have been identified as a crucial resource for historical research on the publishing landscape, but using them requires addressing challenges of data quality, completeness, and interpretation. We call this approach *bibliographic data science*. In this article, we briefly assess the development of book formats and the vernacularization process in early modern Europe. The work undertaken paves the way for more extensive integration of library catalogs to map the history of the book.

ARTICLE HISTORY

Received July 2018
Revised September 2018
Accepted October 2018

KEYWORDS


National bibliography; data ecosystem; publishing history; digital humanities; open science

Introduction

Library catalogs are essential tools in information science, and their utilization has been greatly advanced by digitalization.¹ The need to manage and organize the ever-increasing body of digital information has motivated the development of new concepts and technologies, such as Linked Data, which was first introduced some 20 years ago and has been on the agenda of most National Libraries since then. During the last decade, the concept of Linked Open Data (LOD) emerged to emphasize the importance of open licensing of the data resources. Metadata collections of published material that different libraries hold are particularly suitable for interlinking and enriching with different semantic layers.² LOD represents a crucial step in taking full advantage of digital resources through the integration of web sources and open, reusable metadata and its enrichment.³

National bibliographies have been traditionally used as a tool for information retrieval. This article demonstrates our quantitative approach to book history, where bibliographic collections are considered as research material, rather than a mere retrieval tool. A key feature in this work is

CONTACT Mikko Tolonen  mikko.tolonen@helsinki.fi  Helsinki Computational History Group, Department of Digital Humanities, University of Helsinki, P.O. Box 24 (Unioninkatu 40), FIN-00014

 Supplemental Material for this article can be accessed [here](#).

Published with license by Taylor & Francis © Leo Lahti, Jani Marjanen, Hege Roivainen and Mikko Tolonen.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

that whereas the analysis of full texts has drawn considerable attention in digital humanities, in our analysis metadata collections form the primary target.⁴ This article relates closely to the data management efforts in National Libraries as it claims that it is extremely important that we can rely on data quality and completeness in order to make robust statistical claims. Thus, even when it is certain that no cumulative integrated catalog of bibliographic data will be perfect and free from errors, we argue that a metadata collection can often be sufficiently representative of important trends in the history of the book and knowledge production. This hypothesis comes with substantial research potential but it is yet to be systematically explored and tested. Related earlier studies include work in analytical bibliography and book history that have produced highly interesting interpretations by charting long-term developments in the history in books,⁵ or have at least discussed the different opportunities associated with national bibliographies.⁶ The use of bibliographic metadata as a research object has, however, proven to be challenging as obtaining valid conclusions critically depends not only on the overall understanding of the historical context but also on technical issues of data quality and completeness. Subsequently, research cases that build on quantitative analysis of these data collections, have remained few.

We have started to develop novel ways of addressing these needs by algorithmically harmonizing and integrating different sources of bibliographic metadata maintained by the research libraries. We call this approach *bibliographic data science (BDS)*. It is specifically targeted at enabling the use of bibliographic metadata as a research object, deriving from the more generic paradigms of open science and data science.⁷ We propose that large-scale, automated harmonization efforts can enhance the overall reliability and commensurability between independently maintained metadata collections, thus complementing LOD and other technologies that primarily focus on data management and distribution. Hence, bibliographic data science aims to fill an important gap in the field as it is commonly observed that bibliographic metadata has high amounts of inaccurate entries, collection biases, and missing information. We aim to show how many of these issues can be overcome, so that large-scale quantitative analysis of bibliographic metadata becomes more reliable, by turning to two historical research cases: the rise of the octavo format in printing in Europe and the breakthrough of vernacular languages in public discourse.

Our analysis covers the overall publishing landscape in the period c.1500–1800 based on four large bibliographies. Thus, our analysis allows us to assess publishing activity beyond what is accessible by the use of individual national bibliographies alone, as we have recently suggested.⁸ We have extensively harmonized selected metadata fields of the Finnish and

Swedish National Bibliographies (FNB and SNB, respectively), the English Short-Title Catalogue (ESTC), and the Heritage of the Printed Book database (HPBD). Altogether, these four bibliographies cover over 6 million entries of print products in Europe and elsewhere, and 2.64 million harmonized entries from the investigated period (1500–1800), ranging from the 16,365 entries in the FNB to 2.1 million entries in HPBD, which is a compilation of 45 smaller, mostly national, bibliographies⁹ (Table 1).

Bibliographic data science shifts emphasis from data quantity and management toward data quality and statistical analysis, and has potential for wider implementation in related studies and on other bibliographic metadata collections of which there is certainly no shortage in the Galleries, Libraries, Archives, and Museums (GLAM) sector. Our work indicates that whereas national bibliographies have essentially been about mapping the local canon of publishing, integrating data across borders should be managed in a systematic way that can take into account specific local circumstances. Although print culture has obviously been tied to the nation and its culture, it reflects broader Europe-wide cultural processes that deserve to be analyzed. Integrating data across the borders set by national bibliographies helps us to get at the wider processes and trends and, eventually, to overcome the national view in analyzing the past.

Bibliographic data science

For various reasons, it is important to comprehend that supporting quantitative, data-intensive research is not the original or intended goal of analytical bibliography. Primary motivation for cataloging has been to preserve as much information of the original document and its physical creation as possible. This includes potential errors caused by the printer.¹⁰ If, for instance, a place name is wrongly spelled on the title page, for cataloging purposes it is relevant also to preserve that misspelling. For anyone desiring to work on quantitative approach to bibliographic metadata, this is a crucial point to understand and respect. Moreover, the contents in bibliographic metadata collections are the products of at least three multi-layered historical processes. First, the digitization of traditional card catalogs may have meant an exclusion of material that was regarded as less important or covered elsewhere. Second, the collection of early national bibliographies has in general been based on a collection of existing bibliographies that were originally collected for other purposes.¹¹ Naturally, the national bibliographies have not been able to include everything published, albeit the effort toward completeness has been remarkable in many cases. Third, the records reflect different historical practices of printing, publishing, and cataloging (with respect to variant states of editions, for example).

In 18th-century Sweden, for instance, printing laws and decrees formed a crucial part of political discourse and this was of great economic value to the book industry,¹² whereas in Britain this was the case to a much lesser degree. Such practices are noticeable in the bibliographic metadata collections, but they tell us more about the printing industry, not necessarily about other social and political phenomena, such as language relations, that we might want to study through the data. Any historically interested study using national bibliographies must therefore be attentive to these historical layers contained in the data in order to propose reasonable interpretations of quantitative data analysis. Our work builds on traditional bibliographic research, and we are using established definitions of bibliographic concepts where possible.¹³

Available bibliographic metadata is thus seldom readily amenable to quantitative analysis. Biases, inaccuracies, and gaps hinder productive research use of bibliographic metadata collections. Varying standards and languages pose challenges for data integration. Our use of the term bibliographic data science implies that bibliographic data is viewed as quantitative research material, and systematic efforts on our part are carried out to facilitate this by ensuring data reliability and completeness. In this work, we focus on a few selected fields, namely publication time and place, language, and physical dimensions. Our data harmonization follows similar principles and largely identical algorithms across all metadata collections. We have removed spelling errors, disambiguated and standardized terms, augmented missing values, and developed custom algorithms that can convert the raw MARC notation to numerical page count estimates, for instance.¹⁴ We have also added derived fields, such as *print area*, which quantifies the paper consumption in sheets for a unique copy of a document; the combined print area across different documents in a given time period can be used to quantify the breadth of printing activity. Moreover, we have used external data sources on authors, publishers, and places to enrich and verify bibliographic information. Automation, scalability, and quality control are critical, as the data collections may contain information on millions of documents. Hence, we have incorporated best practices and tools from data science, such as software libraries, unit tests, tidy data and reproducible workflows. Bibliographic data science is based on an iterative process where improved understanding often leads to enhancements in data harmonization and validation that can be incorporated in the automated processing steps.

Ideally, such harmonization and validation efforts are fully transparent both in terms of data and source code.¹⁵ The cumulative research process has equipped us with a vast body of methods that support research use of bibliographic metadata collections. We are sharing our algorithms for

bibliographic data science through the *bibliographica* R package.¹⁶ In contrast to code availability, many of the most comprehensive bibliographic metadata collections are not yet generally available as open data, however, and they may be difficult to obtain even for research purposes. The lack of open data availability forms a major bottleneck for transparent and collaborative development of bibliographic data science. This might be gradually changing, however. The National Library of Finland, for instance, recently made available the complete MARC entries of the FNB¹⁷ under an open data license allowing modification, reuse, and sharing of derivative versions. As we demonstrate with the FNB collection, open data availability enables the sharing of a reproducible workflow from raw data to harmonization and analysis. We share the harmonized version prepared and used in this study; it is openly available and linked from Helsinki Computational History Group website,¹⁸ and can be further verified, investigated, and enriched by others. The harmonized data sets can be further integrated and converted into Linked Open Data and other popular formats in order to utilize the vast pool of existing software tools. As a next step, we are planning to incorporate our validated harmonization algorithms in the Linked Open Data Release of the FNB. Combining large-scale harmonization with existing data management infrastructures could open up new doors for research on national bibliographies.

The HPBD catalog is a compilation, and incorporates parts of the other catalog.¹⁹ In summary, the HPBD contains 19,400 records from FNB (before 1827), and c. 56,000 records from SNB (1600–1800). Hence, the HPBD potentially covers the complete FNB and SNB (Table 1). However, this is not likely to introduce major bias in the current analysis as the smaller FNB and SNB catalogs form a negligible fraction ($\leq 2\%$) of the HPBD. The British Library ESTC collection is not mentioned by name in HPBD, but it is mentioned that HPBD includes 55,400 records from Incunable Short-Title Catalogue and Books printed in the German-speaking countries and of German books printed in other countries (1601–1700) from the British Library; this suggests that the overlap between HPBD and ESTC in the investigated period is at most 0.5% since in ESTC we have identified 1054 Incunables, and 321 German books from German speaking regions (Germany, Austria, Switzerland) printed before 1701.

Data harmonization and management is only the starting point for analysis, albeit an important one. In addition to improving the overall data quality and hence the overall value of LOD and other data retrieval infrastructures, the harmonization enables statistical analysis of the complete metadata collection with scientific programming environments such as R²⁰ or Python²¹, which provide advanced tools for modern data analysis and statistical inference. Whereas large portions of data analysis can be

automated, efficient, and reliable research use requires collaboration between traditionally distinct disciplines, such as history, informatics, linguistics, and data science. Finding the right combination of expertise may be challenging.

Language and format of early modern publications

The hand-press period is particularly fruitful for quantitative research on books because there were remarkably few changes in printing technology from 1450 to approximately the 1830s. It has been famously claimed that Gutenberg himself would have been able to operate a printing press in late 18th-century London since it would have been so similar to the one found in mid-15th-century Mainz. As revolutionary as the movable type printing press was for early modern culture and economy in general, it is a good fortune for our aspirations to understand the development of early modern publishing that there were no game-changing technological innovations for the next 400 years or so after Gutenberg's time.²²

In our research on different bibliographic metadata collections we have come to realize that the relatively stable nature of printing opens up different avenues for cross-European research. For example, we can estimate the long-term development of book formats in some detail across Europe, which in turn is significant for understanding the relevance of printing for the changes in public communication. This is why for this article we have developed two Europe-wide bibliographical metadata cases to analyze the rise of octavo format and process of vernacularization in the early modern period. This tests also the metadata collections in their different levels of data harmonization and respective levels of historical representativity. Both of these research cases represent large-scale Europe-wide transformations that took place predominantly during the hand-press era, but an inspection of them through several metadata collections and by zooming in and out in the material show intriguing variety in the publication profiles of European cities. The cases also make it possible to discuss how the employed methodology, varying levels of data harmonization, and gaps in data affect the analyses, thus paving the way for new research and guidelines for future data integration in this field.

The rise of octavo in the Enlightenment period

The general trend in the metadata collections that we have studied is that the octavo format supersedes other printing formats during the 18th century.²³ This can be measured by looking at a simple title count of documents published in different formats, or by studying the paper consumed

in different documents. We have chosen the latter to better account for books of different sizes and lengths. In this article, we use print area, which quantifies the amount of sheets used for unique copies of titles. Earlier we have also studied total paper consumption, which additionally takes the possibly variable print run estimates into account.

When we examine the publishing trends of book formats in the HPBD, we notice that on a general European level the rise of the octavo format is particularly strong during the 18th century, and further supported by the ESTC and SNB ([Figure 1](#)) where octavo is not only the fastest gainer of the market, but also holds the largest share of the print area by the end of the 18th century. If we look at particular places with respect to octavo share in HPBD, a striking feature is the octavo share in German cities of Frankfurt, Leipzig, Halle, and Berlin ([Supplementary Figure 1](#)). The manner in which folio drops and octavo rises in German soil during the 18th century suggests that the octavo format was the high rising star of the Enlightenment.

Among this type of general Europe-wide trends, there are of course local differences, and for example in Turku ([Supplementary Figure 1](#)), and Finland that was part of Sweden at the time, the rise of octavo comes later than in Sweden in general. This was due to the fact that the main part of the documents printed in Finland were official documents, pamphlets, and theses. If we look at the share of the different formats in Turku, another way of saying this would be that printing in Turku only takes off in the later 18th century whereas in Stockholm the hand press printing industry seems to have reached a different level of maturity earlier ([Supplementary Figure 1](#)). The simplest explanation for the success of the octavo format is that it was particularly suited for smaller books that could be carried around and read practically anywhere, whereas the quarto (and folio) were more commonly used in governmental and academic documents; pamphlets and in larger books alike.²⁴ We have analyzed the relevance of the rise of octavo with respect to book printing in the case of “history” publishing earlier.²⁵ Of course, larger formats in book printing carried certain prestige also in the 18th century, even when reading started to be partly removed from stately mansion libraries and the price of the book turned out to be a decisive factor for dissemination of ideas.²⁶ When considering quarto and octavo publications, it is quite telling that David Hume (1711–1776) wanted his *History of England* to be printed in quarto-sized fine-paper six-volume set in late 1760s (as it had appeared earlier), but the editions that were actually published after 1767 until Hume’s death (including the 1778 posthumous edition) are octavo editions in eight volumes. The octavo editions might have lacked the exclusivity and finesse of heavier tomes with large margins that connoisseurs preferred for aesthetic reasons, but it was particularly the cheaper and smaller formats, octavo and

duodecimo, that changed the nature and relevance of printing and reading in the later part of the 18th century.

We have included one union catalog, ESTC, in this study. It is evident that ESTC is not complete in the sense that it would include all the recorded documents in different libraries. Going through some of the records of larger repositories, such as the National Library of Scotland, quickly reveals that their collections include at least dozens of documents not yet recorded in the cumulative ESTC. This, however, is not a problem for our analysis because we are mainly focused on general trends that do not require all the possible records in order to be reliable.²⁷ One particularly interesting feature of the ESTC is the high proportion of duodecimo documents (Figure 1). At the end of the 18th century, duodecimo in the ESTC reaches the same level as the fast-declining folio. Compared to the proportions of gatherings in the HPBD, for example, this is a highly noticeable feature. In the SNB, folio is at the same level as duodecimo, but the share of the total volume is much lower than in the ESTC. This can be largely explained by differences in the printing costs and because the market responds to the demand of cheaper reads.²⁸ If we analyze different cities based on book format proportions, we realise that it was especially places in North America (such as Boston and Philadelphia), Ireland, and Scotland (Dublin, Edinburgh, and specifically Glasgow) where the duodecimo format has the highest share of the print area (Supplementary Figure 1). Interestingly, in London, folio seems to keep its relatively high share even in the latter part of the 18th century. Also in the traditional university towns of Oxford and Cambridge duodecimo does not rise to the two most common formats in the later 18th century, which is noteworthy in the Anglo-American context.

The most complex data set that we used in this article is the HPBD.²⁹ This is not an integrated metadata unit (such as the ESTC for example), but rather a collection of various bibliographic collections with varying amounts of data with issues of duplicates and the like. Thus, all the analysis of the HPBD need to be executed with additional caution, although we have validated our key observations by ensuring that similar trends are to be found in the other metadata collections that we used. Thus, we can rely on the general trends that are apparent in the HPBD. However, the more specific the analysis becomes, the more careful we need to be. One general feature of the HPBD when it comes to the question of format, along with the earlier noted sharp rise of the octavo, is the relatively large proportion of folio books (Figure 1). It is worth noting that with respect to HPBD, the folio format keeps a fairly large share of the total print area of published documents until mid-18th century. We may notice a similar trend also in ESTC in Figure 1, whereas in SNB folio seems to have been on a sharper drop for a longer

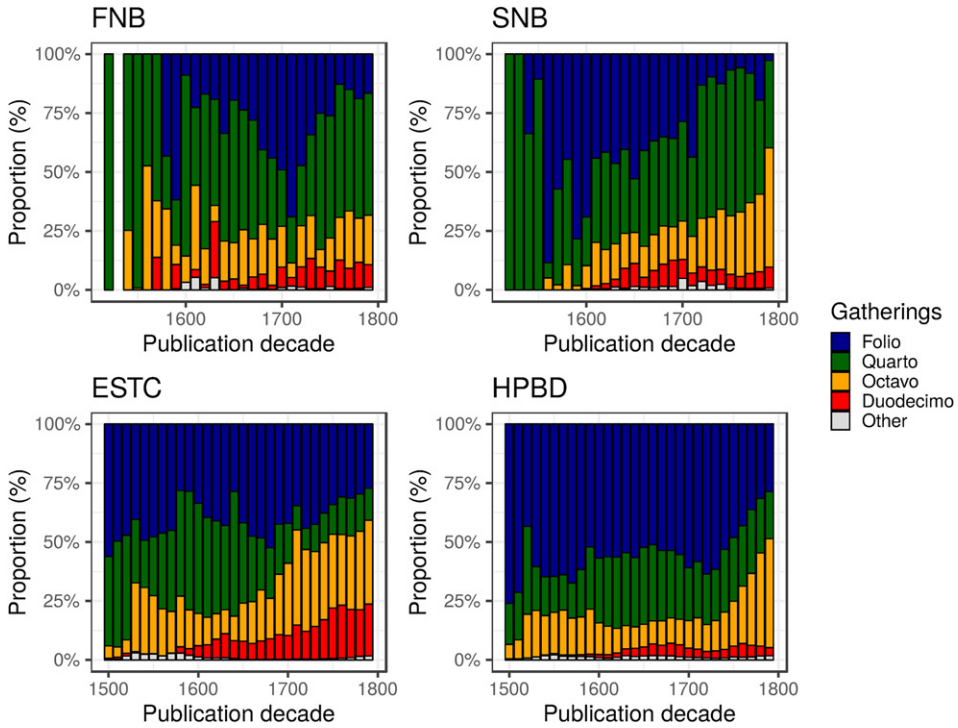


Figure 1. Annual relative print area for common book formats.

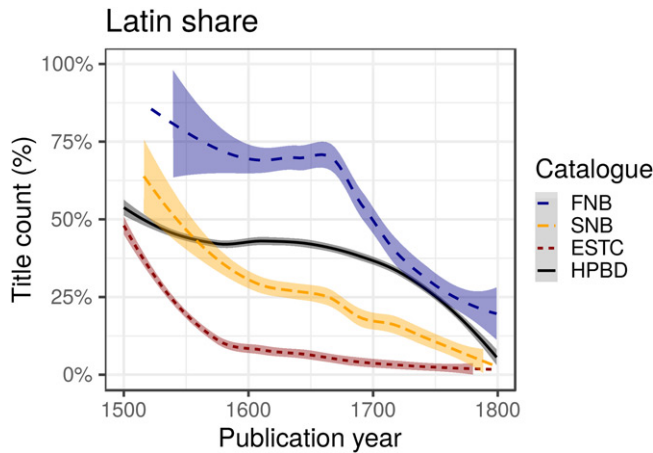


Figure 2. Annual title count proportion of books with Latin as the primary (or only) language.

period of time. Of particular places in HPBD, the large share of folio in Madrid and Brussels ([Supplementary Figure 1](#)) catches one's eye.

The dominant document format in the 17th century together with folio was quarto throughout Europe. There is an unusual peak during the civil war era in ESTC caused by the Thomason Tracts.³⁰ This means that

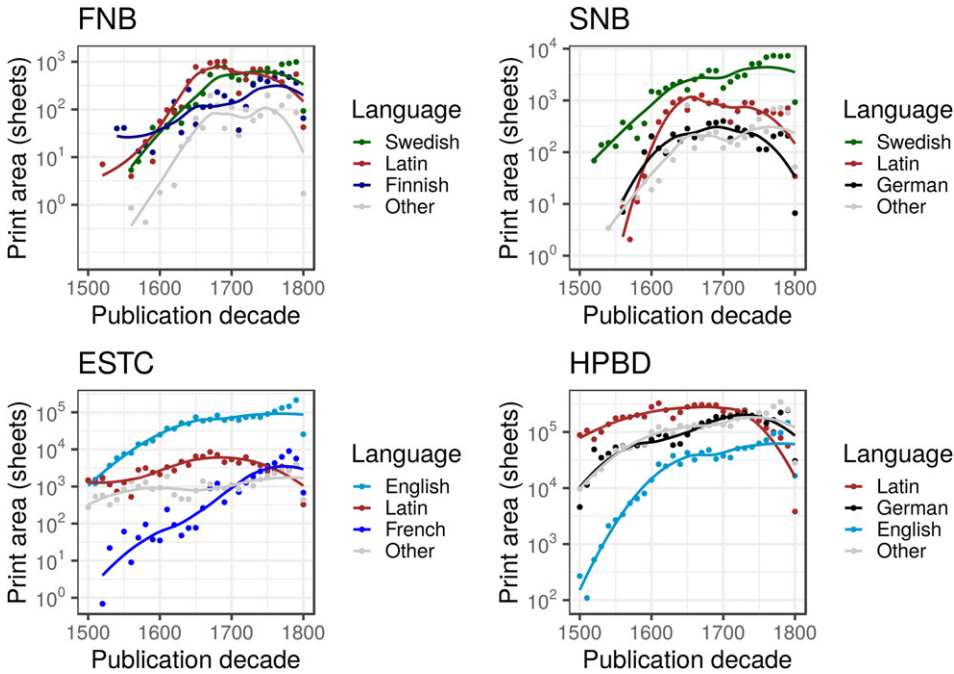


Figure 3. Changes in print area over time. The three most common languages from each catalogue are included.

Table 1. Comparison of the four bibliographic metadata collections used in this study.

Catalogue	Earliest Record	Records 1500-1800 (N)	Language available	Publication place available	Page count available	Gatherings available
FNB	1488	16365	100.0%	93.9%	99.9%	98.3%
SNB	1457	46764	100.0%	95.0%	99.9%	84.8%
ESTC	1473	479790	100.0%	99.4%	99.9%	97.0%
HPBD	1446	2095628	100.00%	86.7%	99.5%	45.3%

because of the cataloging rules of including different variants in ESTC, bookseller George Thomason was able to gather so many of these with respect to civil war pamphlets that there is a noticeable statistical peak caused by them. This needs to be noted, but it does not change the overall general trend.³¹

The quarto format was, as said earlier, the common document format for pamphlets and other shorter pieces. When we look at the HPBD (Figure 1) we see that quarto's share is fairly constant throughout the early modern period. In the ESTC, however, there has been a declining curve since the second half of the seventeenth century. This is because of the more rapid increase of other formats. In the ESTC the quarto format does not decline in absolute numbers, but like all other book formats, its absolute numbers are rising in the 18th century. It is also interesting to notice that there seems to be a correlation between the document language and format.

Comparison of documents published in English, Latin, and other languages in London ([Supplementary Figure 3](#)) suggests that especially duo-decimo was the preferred format for books printed in other languages than English and Latin, whereas octavo was the one used proportionally more in Latin books than others. Especially the small share of folio documents in Latin is interesting. Also, the quarto share of Latin in this respect in London is noteworthy.

Vernacularization in Europe, 1500–1800

Vernacularization refers to a historical transformation in local language relations. Multilingual systems in which one language (in Europe often Latin) was reserved for learned communication, whereas local vernacular languages used in everyday communication started to erode and local languages gained increased prominence. They were made into vehicles for discussing politics, science and culture. This process happened at different speeds in different parts of Europe. Judging from today's teleological perspective, vernaculars such as English and French gained prominence already in the 1600s, whereas for the German and Swedish languages this development happened in the 18th and 19th centuries. For many smaller languages in Europe, such as Finnish or Czech, this development happened in conjunction to nation building in the latter half of the 19th century. Ultimately, vernacularization is an open-ended process. For many potentially vernacularizeable languages the transformation never took place and a similar process could potentially take place also in the future as language relations are in a constant flux. The dominance of English today in many parts of Europe, in a sense, marks a reversed transformation. Linguists and historians have from various perspectives paid attention to vernacularization as a process,³² but this article takes a novel approach by investigating metadata collections that contain millions of titles and related bibliographic information and thus provide a previously unexplored source to trace how the process of vernacularization materialized in concrete publications.

While language relations differ considerably all over Europe, there is one measure that paints a picture of vernacularization as a general trend in European publishing: the share of publications in Latin. All of our four metadata collections show an indisputable declining trend in the share of publications in Latin in the period 1500–1800, but there are noticeable differences to the timing and proportions of the transformation, which are partly explained by historical trajectories mirrored in the data but also by the composition of the data itself. The HPBD ([Figure 2](#)) provides the geographically broadest overview of the decline of Latin, but as a data set it includes most gaps and uncertainties. Nevertheless, in the HPBD the

decline of Latin in the 18th century is most rapid and it happens later than for the ESTC and SNB (Figure 2). This may be a result of the composition of the database with many metadata collections being predominantly focused on the 18th century. The earlier decline of Latin in Britain corresponds with our previous knowledge of the early establishment of English as the main language of high-level communication. Well-known symbols for using English such as Shakespeare and the Royal Society anticipate this,³³ but once the comparison based on national bibliographies can be brought to a more reliable level, we can provide a statistically more accurate picture of this. The available data does nevertheless suggest that the decline of Latin in Britain is more drastic than it has been previously anticipated.

The SNB and FNB allow us to zoom in and look at the Swedish case more closely and compare the different properties of the bibliographies. While the SNB portrays the general trend for the Swedish realm, it is also clear that Stockholm as a publication center dominates the image (Figure 3). Looking at the FNB, which consists mostly of publications from Turku (Åbo), one of the four university towns in the realm (excluding Tartu), shows that the distinct publication profile of university towns are sometimes hidden under the national average. Still, also in Turku, we find a concrete decline in the share of Latin publications, but the decline was definitely later although the Academy in Turku has been described as one of the most utility-oriented universities in the Swedish realm and thus also most prone to use Swedish in academic texts.³⁴ One special feature with the FNB has to do with the different roles of Swedish and Finnish as languages. While Swedish became a stronger candidate for academic publications, Finnish emerged as a written language especially in shorter religious and economic texts. Vernacularization was in this case not a process involving two languages, but three.

Keeping in mind the uncertainties relating to the HPBD, an inspection of smaller university towns suggests that this is a wider trend. The university town, the capital, and the commercial centers had different linguistic publication profiles and vernacularization as a process happened in different phases. An analysis of languages used in publications from Cambridge, Oxford, Leiden, and Göttingen (Supplementary Figure 2) shows how Latin lingered on, but also in these cases, like in Turku the local languages did gain a much more prominent position by the end of the 18th century. Compared to the absolute publishing centers in Europe, Paris, and London, the development happens really late. Interestingly, the metadata collections tell us about national trends, such as an early decline of Latin or competing vernaculars, but when viewed in comparison we can also see patterns that cross national boundaries, such as different types of publishing milieus in

commercial towns, university towns or capital cities. All of Europe had a cultural debt to sources from Antiquity, but this debt materialized differently in the places that were almost self-sufficient culturally (like Paris and London) or the university towns that embodied learning by attaching themselves to Latin traditions.³⁵

Since both vernacularization and the rise of octavo seem to be inherently related to a modernization of public discourse, reading, and writing, a final question is then if the change in the popularity of formats in the 16th and particularly 18th centuries is related to the shifts in language in the same period. It seems that there is no simple answer to this. Quite naturally, in all of the studied metadata collections, the vernacular languages obtain a growing share of published books in the octavo format ([Supplementary Figure 3](#)). A closer look at cities with different publication profiles shows that the matter was more complicated. For the ESTC the share of octavo books is for most cities higher in Latin books than for English books ([Supplementary Figure 4](#)). Also in the SNB, both Latin and Swedish books tend to navigate towards smaller formats at the end of the 18th century, but the HPBD's records for German cities point at the octavo format being used more often in German-language books than in Latin books ([Supplementary Figure 4](#)). While there is not a clear correlation between language and format, the analysis of format nonetheless helps to qualify earlier research. Henrik Horstbøll has shown that the octavo format was particularly popular in Denmark with small histories that stood for a leisurely reading,³⁶ but by looking at a much bigger sample, it is clear that the octavo format became more popular in other genres as well, including books published in Latin in university towns. Additional data and content analysis will in the future allow to look more closely at how genre, language, and format relate to one another, and to what extent the rise of smaller formats and different languages reflect the emergence of new genres.

Discussion

This article has sought to demonstrate that something as seemingly trivial as document sizes and language of titles can have a crucial role when considering the emergence of public sphere in early modern Europe. The relationship between reading habits and broadly circulated written documents in the Enlightenment period can be looked at differently when we can learn more about the relevance of octavo-sized book and the rise of local written languages in Europe during the 18th century. For a better and more reliable understanding of these processes we have developed and used tools of bibliographic data science.

Our work is part of the emerging trend towards the utilization of large digital data resources in publishing history.³⁷ Many of the problems relating to scalable data processing and interpretation were similar to the ones we have encountered in the context of bibliographic metadata collections.

We have investigated four different types of bibliographic metadata collections (FNB, SNB, ESTC, and HPBD). As similar datasets national bibliographies are not only about mapping the national traditions of publishing, but can also be studied comparatively and ultimately be integrated across borders to help to overcome a national perspective in analyzing the past.

The power of a large-scale statistical approach is that broad patterns in knowledge production are often overwhelmingly clear, despite occasional inaccuracies and collection biases in individual data sets as we have shown. Already the HPBD can be used to assess some general trends in publishing history although it does not compete in data reliability and level of harmonization with the other bibliographic metadata collections that we prepared for this study. Unlike for the other collections, we did not customize the harmonization process for HPBD and this should be hence considered preliminary, the correspondence of the observed patterns between this and the other collections demonstrates the scalability of our approach. This is exemplified by our key observations on vernacularization and the rise of the octavo, which are supported by similar trends across the four bibliographic metadata collections that we have assessed. For a more detailed comparison across European cities, further harmonization and augmentation of bibliographic metadata collections are needed.³⁸ Integration of collections demands further work in detecting duplicates, different editions and translations across catalogs in a reliable way. Our present work provides a starting point and the initial guidelines for more extensive analysis and integration.

Bibliographic data science derives from the already established field of data science. It associates this general paradigm specifically with quantitative analysis of bibliographic data collections and related information sources. While having a specific scope, BDS is opening up pragmatically oriented and substantial new research opportunities in this area, as we have aimed to demonstrate.

Our future work envisions continued harmonization and data integration for the HPBD and expanding the study to cover public communication more broadly. As we have extracted and harmonized publisher information from imprints from ESTC and FNB,³⁹ it is possible to connect that data to full-text collections such as the ECCO. Our vision also includes studying how the materiality of printing is related to developments in newspapers.⁴⁰ We consider the material developments within the printing industry as

crucial ingredients in the emergence of new types of public communication that transformed Europe in the 18th century.

Our current harmonization strategies are based on manually implemented rules for data processing. Future developments can take further advantage of established machine learning techniques in order to reduce the need for human input and improving the overall scalability of data harmonization. Modeling the emergence of the publishing landscape across Europe could also borrow spatio-temporal analysis methods from ecology and related fields. When combined with a proper quality control, such quantitative, data-driven approaches can have potential for wider implementation in related studies in the digital humanities. Moreover, digitalization has provided new opportunities for open sharing of research data and analysis methods. Taking full advantage of these developments can support collaborative and cumulative research use of bibliographic collections.

Conclusion

We have conceptualized a new approach, bibliographic data science, to expand the research potential of bibliographic cataloging and classification. Whereas national bibliographies can provide comprehensive quantitative insights to the overall historical dynamics of the evolving publishing landscape across time and geography, we have encountered specific and largely overlooked challenges in using bibliographic metadata collections for historical research. Biases, gaps, and inaccuracies in data collection may remarkably hinder productive research use of the bibliographies, and drawing valid conclusions critically depends on efficient and reliable harmonization and augmentation of the raw entries. Here, we have overcome some of these challenges by specifically tailored open data analytical ecosystems that facilitate robust statistical research use of bibliographic collections. This approach has potential for wider implementation in related studies and bibliographies, and provides guidelines for more extensive integration of national collections, thus moving towards a more precise view of print culture beyond the confines of national bibliographies.

Supplementary Material

The source code and harmonized version of the Finnish national bibliography Fennica (FNB) used in this study, based on the original open MARC records published by National Library of Finland, is available through Helsinki Computational History Group (COMHIS) website.⁴¹

Acknowledgments

We are grateful to the National Library of Finland, the National Library of Sweden, the British Library, and CERL for providing the bibliographies for use in this research, and for the members of Helsinki Computational History Group for supporting this work.

Funding

This work was supported by the Academy of Finland under Grant 293316.

ORCID

Leo Lahti  <http://orcid.org/0000-0001-5537-637X>

Jani Marjanen  <http://orcid.org/0000-0002-3085-4862>

Hege Roivainen  <http://orcid.org/0000-0002-0489-3278>

Mikko Tolonen  <http://orcid.org/0000-0003-2892-8911>

Notes

1. In the following text when we discuss library catalog or national bibliography data, we refer to it mainly as bibliographic metadata or bibliographic metadata collections for the sake of clarity. This article is meant as a positive contribution to a long-standing discussion described in G. Thomas Tanselle, “Bibliography and Science,” *Studies in Bibliography* 27 (1974): 55–90.
2. Good examples of LOD services among National Libraries are Kansalliskirjasto, *Finto - suomalainen sanasto- ja ontologiapalvelu*, (accessed June 29, 2018), <http://finto.fi/en/> and British Library, *The British National Bibliography as Linked Open Data*, (accessed June 29, 2018), <http://bnb.data.bl.uk/>. About development of LOD in the library sector, see Becky Yoose and Jody Perkins, “The Linked Open Data Landscape in Libraries and Beyond,” *Journal of Library Metadata* 13, nos. 2–3 (2013): 197–211, DOI: <https://doi.org/10.1080/19386389.2013.826075>.
3. Tom Heath and Christian Bizer, “Linked Data: Evolving the Web into a Global Data Space,” *Synthesis Lectures on the Semantic Web: Theory and Technology* 1, no. 1 (2011): 1–136, DOI: <https://doi.org/10.2200/s00334ed1v01y201102wbe001>.
4. Metadata is often considerably more structured, standardized, and smaller in size than full texts, greatly facilitating automated analysis. Moreover, metadata could provide valuable context for interpreting full text collections, as bibliographic metadata is often available for a larger number of documents than digitized full texts are, and in a more standardized format. About our work combining metadata and full text sources, see also the posters by Kanner et al. “OCTAVO – Analysing Early Modern Public Communication”, presented in Digital Humanities at Oxford Summer School 2017; and Ginter et al. “Metadata Analysis and Text Reuse Detection: Reassessing public discourse in Finland through newspapers and journals 1771–1917 in Digital humanities in the Nordic Countries DHN2018 – University of Helsinki, 2018. The posters are available at Helsinki Computational History Group website (<http://comhis.github.io>).

5. Eltjo Buringh and Jan Luiten van Zanden, "Charting the "Rise of the West": Manuscripts and Printed Books in Europe A Long-Term Perspective from the Sixth through Eighteenth Centuries," *The Journal of Economic History* 69, no. 02 (2009): 409, DOI: <https://doi.org/10.1017/s0022050709000837>; Joerg Baten and Jan Luiten van Zanden, "Book production and the onset of modern economic growth," *Journal of Economic Growth* 13, no. 3 (2008): 217–235, DOI: <https://doi.org/10.1007/s10887-008-9031-9>; Maureen Bell and John Barnard, "Provisional Count of STC Titles, 1475–1640," *Publishing History* 31, no. 1 (1992): 47–64; Henrik Horstbøll, *Menigmands medie: det folkelige bogtryk i Danmark 1500-1840: en kulturhistorisk undersøgelse*, Danish humanist texts and studies, volume 19 (Copenhagen: Det Kongelige Bibliotek & Museum Tusculanum, 1999), 01058746.
6. Michael F. Suarez, "Towards a bibliometric analysis of the surviving record 1701–1800," in *The Cambridge History of the Book in Britain*, ed. Michael F. Suarez and Michael L. Turner (Cambridge University Press, 2009), 37–65, DOI: <https://doi.org/10.1017/cho9780521810173.003>; Michael F. Suarez, "Book history from descriptive bibliographies," in *The Cambridge Companion to the History of the Book*, ed. Leslie Howsam (Cambridge University Press, 2014), 199–218, DOI: <https://doi.org/10.1017/cc09781139152242.015>; Alexis Weedon, "The Uses of Quantification," in *A Companion to the History of the Book*, ed. Simon Eliot and Jonathan Rose (London: Blackwell Publishing Ltd., 2008), 33–49, DOI: <https://doi.org/10.1002/9780470690949.ch3>.
7. See Leo Lahti, "Open Data Science," in *Advances in Intelligent Data Analysis XVII*, ed. Wouter Duivesteijn, Arno Siebes, and Antti Ukkonen (Cham: Springer International Publishing, 2018), 31–39.
8. Mikko Tolonen, Leo Lahti, Hege Roivainen, and Jani Marjanen, "A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828," *Historical Methods*, In press, DOI: <https://doi.org/10.1080/01615440.2018.1526657>.
9. <https://www.cerl.org/resources/hpb/content>.
10. For a good discussion of W. W. Greg and Fredson Bowers who largely shaped the field, see also Neil Harris (2004): Definitions of Bibliography, and in Particular of the Variety Called Analytical; <http://ihl.enssib.fr/analyticalbibliography-an-alternative-prospectus/definitions-of-bibliography-and-in-particular-of-the-variety-called-analytical> (accessed on, June 29, 2018).
11. For a related discussion on the Danish National Bibliography, see Horstbøll, *Menigmands medie: det folkelige bogtryk i Danmark 1500-1840: en kulturhistorisk undersøgelse*.
12. Anna-Maria Rimm, "Den kungliga boktryckaren, del 1," *Biblia* 30 (2005): 4–31; Anna-Maria Rimm, "Den kungliga boktryckaren, del 2," *Biblia* 31 (2005): 27–44.
13. For most analytical bibliographical definitions, we rely on Philip Gaskell, *A New Introduction to Bibliography* (New Castle, DE: Oak Knoll Press, 1972, rev. ed. 1995).
14. Field 300a for the page count estimates; see the Library of Congress web document <https://www.loc.gov/marc/bibliographic/> for the full description of the MARC21 format.
15. Andrew Morin, Jennifer Urban, Paul D. Adams, Ian Foster, Andrej Sali, David Baker, and Piotr Sliz, "Shining Light into Black Boxes," *Science* 336, no. 6078 (2012): 159–160, DOI: <https://doi.org/10.1126/science.1218263>.
16. <https://github.com/COMHIS/bibliographica>.
17. <http://data.nationallibrary.fi/>.
18. <https://comhis.github.io/2019> CCQ.
19. For a detailed description, see <https://www.cerl.org/resources/hpb/content>.

20. R Core Team, R: *A Language and Environment for Statistical Computing* (Vienna: R Foundation for Statistical Computing, 2018).
21. Python Software Foundation, *Python Language Reference, version 3.6.5*, 2018.
22. David McKitterick, *Print, Manuscript and the Search for Order, 1450-1830* (Cambridge University Press, 2005) About the relevance of movable type printing press, see Elizabeth L. Eisenstein, *The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early-Modern Europe*, Volumes 1 and 2 in one (New York: Cambridge University Press, 1980); Carlo M. Cipolla, "The Diffusion of Innovations in Early Modern Europe," *Comparative Studies in Society and History* 14, no. 1 (1972): 46–52, <http://www.jstor.org/stable/178059>; Andrew Pettegree, "Centre and Periphery in the European Book World," *Transactions of the Royal Historical Society* 18 (2008): 101–128, <http://www.jstor.org/stable/25593882>. On economic impact of printing press on early modern cities: Jeremiah E. Dittmar, "Information Technology and Economic Change: The Impact of The Printing Press," *The Quarterly Journal of Economics* 126, no. 3 (2011): 1133–1172, DOI: <https://doi.org/10.1093/qje/qjr035>. See also, Anne E.B. Coldiron, *Printers without Borders: Translation and Textuality in the Renaissance* (Cambridge: Cambridge University Press, 2015); Anne E.B. Coldiron, "Public Sphere/Contact Zone: Habermas Early Print, and Verse Translation," *Criticism* 46, no. 2 (2004): 207–222, DOI: <https://doi.org/10.1353/crt.2004.0040>.
23. Henrik Horstbøll has previously studied the relevance of octavo format for Danish publishing in detail based on analogue methods and smaller samples. Our work confirms his findings and further extends the scope by studying a much larger and cross-European data. See Henrik Horstbøll, "In octavo: formater, form og indhold på det litterære marked i 1700-tallets Danmark", in *Bokens materialitet : bokhistoria och bibliografi : bidrag till en konferens anordnad av Nordiskt nätverk för editionsfilologer, September 14-16, 2007*, ed. Mats Malm, Barbro Ståhle Sjönell, and Petra Söderlund (Svenska Vitterhetssamfundet, 2009), 197–223; Henrik Horstbøll, "Bibliografi som boghistorie", *Biblis* 50 (2010): 90–95, ISSN: 1403-3313 and Douglas Galbi (2011): *Books Shifted to Smaller Formats across Centuries of Print* in <https://www.purplemotes.net/2011/08/21/books-shifted-to-smaller-formats-across-centuries-of-print/> (accessed July 1 2018).
24. About relationship between books and pamphlets, see Joad Raymond, *Pamphlets and Pamphleteering in Early Modern Britain*, Cambridge Studies in Early Modern British History (Cambridge: Cambridge University Press, 2003).
25. Mikko Tolonen, Leo Lahti, and Niko Ilomäki, "A Quantitative Study of History in the English Short-Title Catalogue (ESTC), 1470-1800", *Liber quarterly* 25, no. 2 (2015): 87–116, DOI: <https://doi.org/10.18352/lq.10112>.
26. David W. Allan, "Politeness and the Politics of Culture: An Intellectual History of the Eighteenth-Century Subscription Library," *Library & Information History* 29, no. 3 (2013): 159–169, DOI: <https://doi.org/10.1179/1758348913z.00000000036>; David W. Allan, *Making British Culture: English Readers and the Scottish Enlightenment, 1740-1830* (United States: Routledge, 2008); David W. Allan, *A Nation of Readers: The Lending Library in Georgian England* (London: The British Library, 2008); Mark Towsey, *Reading the Scottish Enlightenment* (Leiden: Brill, 2010), DOI: <https://doi.org/10.1163/ej.9789004184329.i-364>.
27. Jean-Baptiste Michel, Yuan K. Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez L. Aiden, "Quantitative

- Analysis of Culture Using Millions of Digitized Books,” *Science* 331, no. 6014 (2010): 176–182, DOI: <https://doi.org/10.1126/science.1199644>.
28. William St. Clair, *The Reading Nation in the Romantic Period* (Cambridge: Cambridge University Press, 2007).
 29. <https://www.cerl.org/resources/hpb/main>.
 30. <https://www.bl.uk/collection-guides/thomason-tracts>.
 31. In the Helsinki Computational History Group we are now working on algorithms for edition level harmonization in which the objective is to analyse the documents based on first editions or particular editions. This will benefit text mining of different full text resources (such as the Eighteenth-Century Collections Online (ECCO) when combined to ESTC).
 32. Charles A. Ferguson, “Diglossia,” *Word* 15, no. 2 (1959): 325–340, DOI: <https://doi.org/10.1080/00437956.1959.11659702>; Joshua A. Fishman, “Bilingualism With and Without Diglossia; Diglossia With and Without Bilingualism,” *Journal of Social Issues* 23, no. 2 (1967): 29–38, DOI: <https://doi.org/10.1111/j.1540-4560.1967.tb00573.x>.
 33. Ryan J. Stark, *Rhetoric Science, and Magic in Seventeenth-century England* (Washington, DC: Catholic University of America Press, 2011), DOI: <https://doi.org/10.2307/j.ctt2850xr>; Peter Dear, “Totius in Verba: Rhetoric and Authority in the Early Royal Society,” *Isis* 76, no. 2 (1985): 145–161, DOI: <https://doi.org/10.1086/353797>; James Livesey, *Civil Society and Empire: Ireland and Scotland in the Eighteenth-Century Atlantic World*, The Lewis Walpole Series in Eighteenth-Century Culture and History (New Haven cT: Yale University Press, 2009).
 34. Bo Lindberg, “Utilism och upplysning – exemplet Åbo,” *Ugglan* 2 (1993): 34–42.
 35. For Latin-vernacular diglossia, see Bo Lindberg, “Den antika skevheten. Politiska ord och begrepp i det tidigmoderna Sverige,” *Filologiskt arkiv* (Stockholm) 45 (2006).
 36. Horstbøll, “In octavo: formater, form og indhold på det litterære marked i 1700-tallets Danmark.”
 37. For instance, the Culturomics project (Michel, Shen, Aiden, Veres, Gray, The Google Books Team, Pickett, Hoiberg, Clancy, Norvig, Orwant, Pinker, Nowak, and Aiden, “Quantitative Analysis of Culture Using Millions of Digitized Books”) analyzed broad historical trends in English language and culture in the period 1800–2000 based on an unforeseen corpus collected from the full text content of over five million digitized books. On difficulties in interpreting the data, see also Elise E. Morse-Gagne, “Culturomics: Statistical Traps Muddy the Data,” *Science* 332, no. 6025 (2011): 35–35, DOI: <https://doi.org/10.1126/science.332.6025.35-b>.
 38. For instance, Buringh and Zanden, “Charting the “Rise of the West”: Manuscripts and Printed Books in Europe, A Long-Term Perspective from the Sixth through Eighteenth Centuries” rely on an earlier version of the HPBD to assess very general trends in printing, but their methods of data harmonization do not allow delving further into detail.
 39. Mikko Tolonen, Leo Lahti, Hege Roivainen, and Niko Ilomäki, “Printing in a Periphery: a Quantitative Study of Finnish Knowledge Production, 1640–1828,” in *Digital Humanities 2016: Conference Abstracts* (Jagiellonian University & Pedagogical University, Kraków, 2016), 383–385, dh2016.adho.org/static/data/170.html.
 40. Jani Marjanen, Ville Vaara, Antti Kanner, Hege Roivainen, Eetu Mäkelä, Leo Lahti, and Mikko Tolonen, “Analysing the Language, Location and Form of Newspapers in Finland, 1771–1910” (DHN2017 Gothenburg, 2017), https://comhis.github.io/assets/files/presentation_analysing-language-location-and-form_Gothenburg2017.pdf.
 41. https://comhis.github.io/2019_CCQ/.